Journal of Ethnopharmacology 139 (2012) 693–694

Contents lists available at SciVerse ScienceDirect

Journal of Ethnopharmacology

journal homepage: www.elsevier.com/locate/jethpharm

Letter to the Editor

Commentary: Regression residual vs. Bayesian analysis of medicinal floras

A recent paper in JEP proposed using Bayesian statistics rather than regression-residual techniques to analyze medicinal floras (Weckerle, 2011). Since the paper directly critiques a method which I developed, and which has been used by myself and many others since, I feel it is appropriate for me to share a critique of the new method.

In this paper, the authors propose to use Bayesian statistics to analyze a regional Italian medicinal flora. The overall approach uses families as the basic unit of analysis, and creates an n-tuple of pairs of numbers, one pair for each family (128 of them in this case), where the numbers are, first, the number of species in the family in the region, and second, the number of those species reportedly used as medicinals. When I first attempted an analysis like this in the 1970s (Moerman, 1979), the obvious first thing to try was percentages. But that did not work well because there were a number of small families with 1, 2 or 3 species, all or some of which were used medicinally, which thereby had very large percentages (1 of 1, or 2 of 2, or 3 of 3, all rated 100% usage). These small but utilized families swamped the much larger families (Asteraceae, Rosaceae, etc.) that provided the bulk of medicinal plants. To control for family size, I decided to use regression analysis to address the issue.

I found this to be very satisfactory for a number of reasons. First, the regression analysis gave a number of species that might be utilized as medicines in each family if they were selected at random. At the time, it was common for people to suggest indeed, that was just how these people had selected their plants: "they just tried anything, and sometimes they got lucky" was the tone of many writers. This analysis of mine, showing some very large families producing the bulk of the medicinal plants, while other really large families (Poaceae, Cyperaceae, Orchidaceae) did not, certainly sank that silly idea.

It is the case that medicinal plant data that I was analyzing do not meet all the criteria of a regression analysis as detailed by Bennett and Husby (2008). But ordinarily, regression analysis is done on a sample drawn from a larger population. In most of the cases where people use such an analysis on medicinal plants, the assumption (more or less legitimate depending on the circumstances) is that the data are not a sample, but a census of all medicinal plants used, and a full flora listing all the plants available to the people being studied. The last edition of my database of native American plant use (Moerman, 1998) was published 13 years ago. Since that time I have paid careful attention to subsequently published instances of native American plant uses; I have in that time found about a dozen instances of utilized plants not mentioned already in the database. Adding a dozen uses to 47,000 items representing 3618 species will not change anything.

But in addition to the predicted number of species, the regression analysis also pointed to the residual; that is, the difference between the predicted value and the actual value. These gave a clear array of numbers showing precisely which families produced more medicinal plants, and which produced less, than randomness would predict.

These results were fascinating and, I believe, unprecedented. They not only answered once and for all the assertion that indigenous peoples' plant use was random, it raised a number of additional new ethnobotanical questions, like, "Why sunflowers, and not grasses," and a range of similar ones. The analysis allowed a fascinating comparison of food plants and drug plants which overlap in curious and unanticipated ways, and pointed out the significance of "apparency" in the selection of medicinals, and many others as well. And they offered a way to compare widely separated peoples' medicinal plant choices.

In addition, and this is perhaps most important, although regression analysis was more complicated than percentages, it was not difficult to explain, and many people who were up until then unaware of the possible value of a mathematical approach to ethnobotanical issues realized this was a useful and interesting way to proceed. The evidence of this is everywhere in ethnobotanical journals where regression analyses are frequently parts of papers from around the world, and in fascinating comparative analyses. For example, the most recent issue of JEP – which appeared in my mailbox two days before I wrote the first version of this commentary – has a fascinating comparison of the medicinal floras of Nepal, New Zealand and the South African Cape (Saslis-Lagoudakis et al., 2011) using regression residual analysis (they also report Bennett and Huseby's binomial values, but do not make much of them).

Why are the authors of the paper under review unhappy with regression residual analysis? They argue that such an analysis is

> "biased towards large plant families. This is because the residual of a relatively small plant family (e.g. $n = 10$) is maximally 10, while a relatively large plant family (e.g. $n = 100$) may get a residual of up to 100. Nonetheless, in the regression analysis all plant families are lumped together pretending that small families may achieve residuals with the same order of magnitude as those of large families. (p. nn)"

This is simply wrong. No family can get a residual equal the size of the family unless the predicted value for the family were zero, and the actual value were 10 (one calculates the residual by subtracting the predicted value from the actual value). The only way the predicted value could be zero would be in a culture which utilized no medicinal plants, in which case the actual value would be zero (in such a culture, the actual values and the predicted values would all be zero). In the Campagna flora, family sizes range from 1 to 253, and residuals vary from 18.4 to −17.9.

Of course large families *may* produce larger numbers of medicinal species than smaller ones (although it is very rare to find large families with more than 20–30% medicinal; that is, the residuals for large families are never more than about 20, not "100" as alleged in

doi:10.1016/j.jep.2011.09.021

Table 1
Comparison of bottom seven families in Weckerle et al.'s Bayesian analysis, and the author's regression analysis.

Bayes	Regression
Fabaceae	Geraniaceae
Amaranthaceae	Amaranthaceae
Geraniaceae	Cyperaceae
Cyperaceae	Fabaceae
Poaceae	Orchidaceae
Caryophyllaceae	Caryophyllaceae
Orchidaceae	Poaceae

the quotation above). But the whole point of the analysis is to show that many very large families may produce no medicinal species at all. Indeed, anyone interested in the smaller families, or particular smaller families, can easily examine them in context and in perspective by eliminating the larger families in any way they prefer (calling them outliers), and repeating the regression residual analysis.

It is indeed the case that in most floras, a relatively small number of relatively large plant families produce the great bulk of medicinal plants. This is not the result of some statistical trick, rather it is the result of generations of human beings studying nature, finding important and helpful plants, and applying them to self or other in the presence of illness, hunger, or the desire for color, beauty, or structure.

There is, then, a very good reason to privilege large families, because that is where the medicinal plants are. To show this, I carried out a standard regression residual analysis of the data presented by Weckerle et al. In their analysis, they look at the top 14 families as being the most productive of medicinal plants. However, those families contain, respectively, 261 species and 109 medicinal species. This is 12% of the overall flora, and 25% of the medicinal flora. In my regression residual analysis of their data, the top 14 families contain, respectively, 822 species and 228 medicinal species. This is 36% of the overall flora, and 54% of the medicinal species. And, not surprisingly, in the regression analysis, the medicinal flora looks remarkably like the medicinal floras of other northern hemisphere cultures (Moerman et al., 1999).

The bottom 7 families are the same in both analyses, with the order slightly varying (Table 1).

This Bayesian analysis does not clearly describe the medicinal flora (although it does a pretty good job describing the part of the flora which does not provide many medicinals)! It dramatically privileges small families and ignores important large ones. Indeed, the correlation between the *percentage of plants* used and the "Inf." shown in Appendix A is 0.804 ("Inf." is the abbreviation for "inferior 95% probability margin"). The correlation of percent used with regression residuals is 0.29. We can learn very nearly as much about this dataset with an analysis of percentages as we can with all these elaborate mathematics (because, like the Bayesian analysis, percentages privilege small families).

The final problem here is the unnecessary complexity of Bayesian analysis. I have studied a lot of mathematics in my life, and I have not a clue what Section 2.4 is about. I know what the assumptions are in a regression residual analysis; but I have not any idea what they are here. The figures are as obscure as the text.

It is the case that even avowed Bayesian aficionados (like Andrew Gelman) have serious questions about the approach. I encourage everyone to read this article, and the debate it generated (Gelman, 2008).

From my perspective, regression analysis is a simple tool which offers anyone the opportunity to analyze an ethnobotanical dataset, and understand what is happening.

To see the spreadsheet which compares the Bayesian, regression, and percentage analyses, click on this link, or paste it into your browser.

http://www-personal.umd.umich.edu/~dmoerman/Campagna.xls

References

Bennett, B.C., Husby, C.E., 2008. Patterns of medicinal plant use: an examination of the Ecuadorian Shuar medicinal flora using contingency table and binomial analyses. Journal of Ethnopharmacology 116, 422–430.
Gelman, A., 2008. Objections to Bayesian statistics. Bayesian Analysis 3, 445–450.
Moerman, D.E., 1979. Symbols and selectivity: a statistical analysis of native American medical ethnobotany. Journal of Ethnopharmacology 1, 111–119.
Moerman, D.E., 1998. Native American Ethnobotany. Timber Press, Portland, OR.
Moerman, D.E., Pemberton, R.W., Kiefer, D., Berlin, B., 1999. A comparative analysis of five medicinal floras. Journal of Ethnobiology 19, 46–67.
Saslis-Lagoudakis, C.H., Williamson, E.M., Savolainen, V., Hawkins, J.A., 2011. Cross-cultural comparison of three medicinal floras and implications for bioprospecting strategies. Journal of Ethnopharmacology 135 (2), 476–487.
Weckerle, C.S., Cabras, C., Eugenia Castellanos, M., Leonti, M., 2011. Quantitative methods in ethnobotany and ethnopharmacology: considering the overall flora – hypothesis testing for over- and underused plant families with the Bayesian approach. Journal of Ethnopharmacology 137 (1), 837–843.

Daniel E. Moerman*
University of Michigan-Dearborn, Behavioral Science Department, 6515 Cherry Hill Rd, Ypsilanti, MI 48198, United States

* Tel.: +1 734 4833283; fax: +1 734 4801908.
E-mail address: dmoerman@umich.edu

9 September 2011
Available online 29 September 2011